Statistical Inference Course Project (Part 1)

BELMOUIDI Mohamed

27 December, 2024

Course Project

Statistical Inference Course Project

Overview

The Central Limit Theorem states that if you take sufficiently large random samples (n > 30) from a population with mean μ and standard deviation σ, the distribution of sample means will be approximately normal, centered around μ, regardless of the population's shape.

This project explores the Central Limit Theorem using the exponential distribution in R. The theoretical normal distribution is compared to the distribution of sample means from the exponential distribution.

Simulations

Perform 1000 simulations, each with 40 samples of an exponential distribution. The 40 samples will be used to calculate the arithmetic mean and variance and then compared to the theoretical estimates.

To make the data reproducible, a seed will be set. Also, set the control parameters $\lambda = 0.2$ (the rate) and $n = 40$ (number of samples).

```r
# set seed for reproducability
set.seed(062000)

# set sampling values:
lambda <- 0.2              # rate parameter
n <- 40                    # number of samples (exponentials) in each simulation
numSimulations <- 1000     # number of simulations

# simulate the population
simMeans <- data.frame(expMean = sapply(1 : numSimulations, function(x) {mean(rexp(n, lambda))}))
```

Sample Mean vs. Theoretical Mean

The Central Limit Theorem states that sample means will be approximately normally distributed with a mean equal to the population mean μ. For an exponential distribution, the theoretical mean is $\frac{1}{\lambda}$. In this simulation, we compare the sample mean with the theoretical mean.

Calculate the sample and theoretical means across 1000 simulations of 40 samples from an exponential distribution with $\lambda = 0.2$.

```r
# calculate sample mean and theoretical mean
sampleMean <- mean(simMeans$expMean)
theoMean <- 1/lambda
compMeans <- data.frame(sampleMean, theoMean)
names(compMeans) <- c("Sample Mean", "Theoretical Mean")
print(compMeans)
```

```
##   Sample Mean Theoretical Mean
## 1    4.950877                5
```

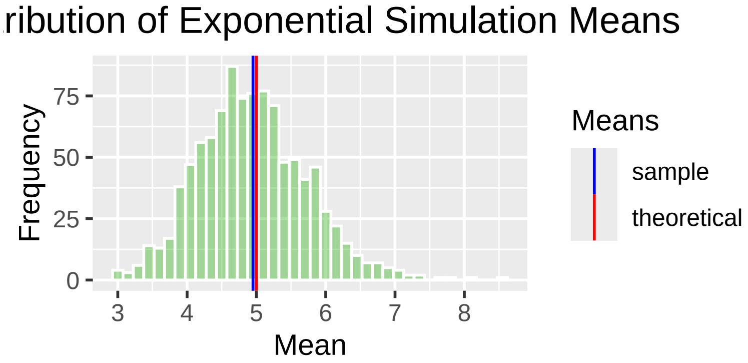
As part of the data analysis, also perform a one sample t-test to check the 95% confidence interval for the sample mean.

```r
t.test(simMeans$expMean, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  simMeans$expMean
## t = 198.48, df = 999, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  4.901927 4.999826
## sample estimates:
## mean of x
##  4.950877
```

Plot Distribution

Display a histogram to show the averages of the 40 exponentials over 1000 simulations. Include the sample mean and theoretical mean for comparison.



The sample mean came out to be 4.9508767, while the theoretical mean is 5. As shown in the chart, the mean of the sample means (blue vertical line) is close to the theoretical mean (red vertical line). With a 95% confidence interval, the sampled mean is between 4.9019272 and 4.9998263, which closely match.

Sample Variance vs. Theoretical Variance

In the same manner used to compare the Sample Mean and Theoretical Mean, the Sample Variance will be compared to the Theoretical Variance.

The theoretical variance is $\frac{(\frac{1}{\lambda})^2}{n}$.

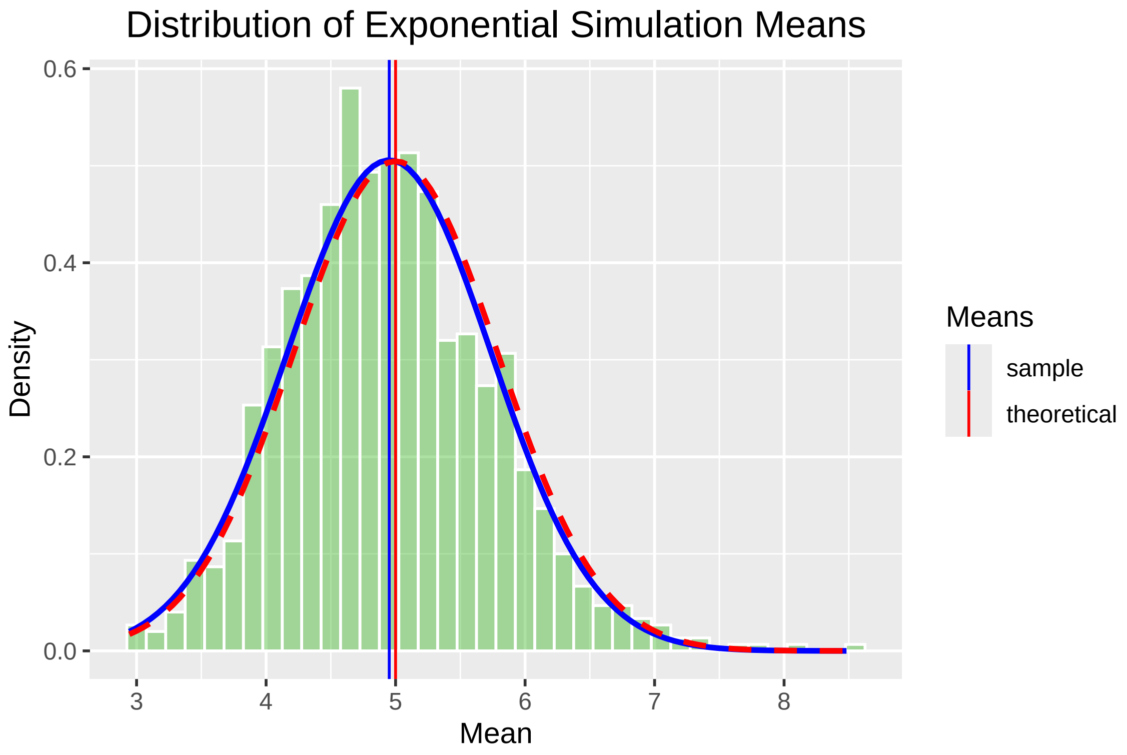
```
# calculate sample variance and theoretical variance
sampleVariance <- var(simMeans$expMean)
theoVariance <- ((1/lambda)^2)/n
compVariance <- data.frame(sampleVariance, theoVariance)
names(compVariance) <- c("Sample Variance", "Theoretical Variance")
print(compVariance)
```

```
##   Sample Variance Theoretical Variance
## 1       0.6222257                0.625
```

The sample variance came out to be 0.6222257 which is very close to the theoretical variance 0.625.

Distribution

Determine whether the exponential distribution is approximately normally distributed about the population mean. According to the Central Limit Theorem, the means of the sample simulations should follow a normal distribution.



As shown in the above plot, the distribution of means of the sampled exponential distribution appear to follow a normal distribution.

The density of the sampled data is shown by the light green bars. The dotted red line represents a normal distribution which is very close to the sample distribution colored in blue.